New algorithm to find a shape of a finite set of points

N. Soukhoroukova [1], J. Ugon
Centre for Informatics and Applied Optimization,
School of Information Technology and Mathematical Sciences,
University of Ballarat, Victoria 3353, Australia
University of Ballarat

Abstract

Very often in data classification problems we have to determine a shape of a finite set of points within datasets. One of the most common approaches to represent such sets is to determine them as collections of several groups of points. The goal of this project is to develop some algorithms to find a shape for each group. Numerical experiments using the Discrete Gradient method have been done. The results are presented.

Introduction

The main goal of data mining is to extract non-trivial information from data (see [1]). Several models (most of them are mathematical based models) have been designed for that purpose (see [2]-[8]). This area has many practical applications as diverse as medical diagnostics, financial decisions, study of experimental data, etc...

With the growth of new technologies, the datasets in the future are going to increase significantly. The datasets we consider in our project are finite sets of N points (observations) in the space $\mathbb{R}^n$. The number N could be more than 100,000. The number n is the number of features we used to describe each observation. The development of fast and accurate techniques to analyze data is thus of great interest.

One problem arising is to divide a dataset into sets of points having common characteristics. These groups could be determined before we start to study the dataset(different classes of observations) or we could do it using some clustering methods (a large variety of methods can be found in [5], [7], [9], [10]).

Once these sets have been obtained, the next step is to study their properties, and to find some common characteristics for all or most of the points within each of these groups. The goal of this project is to develop some algorithms to study geometrical properties of a given finite set of points. We find some objects containing these points. We focus on a special class of geometrical objects, the ellipsoids, their combinations and intersections.

Our algorithm requires the solution of a sequence of optimization problems. We use the Discrete Gradient method to solve these problems. In our study, we apply the new algorithm to study some real-world datasets. We used subsets of features obtained by feature selection methods (see [12] for information), and divided the datasets into groups of points using some clusterization methods (see for example [9]-[11]).

Then we study the shape and the structure of each set. When there are no more than 3 features, the different geometrical objects obtained are graphically represented.

[1]corresponding author nsoukhoroukova@students.ballarat.edu.au

1 Geometrical approximation for a finite set of points

1.1 *Sets and shapes*

Suppose that we have a finite set of points

$$A = \{a_i \in \mathbb{R}^n | i = 1, N\}.$$

Our task is to find a shape containing the set. We would like to construct a geometrical object from a certain class which contains this finite set.

In some approaches developed recently (see for example [11]) the form of the finite set under consideration was a ball. We would like to generalize this proposition and find the form of the finite set in the class of ellipsoids. This ellipsoid can be considered as a continuous approximation of the finite set.

We propose an approach to find ellipsoidal approximations for given sets. In order to describe this algorithm we need some definitions.

Definition 1 An ellipsoid which includes the finite set of points A is called a *shape* of the set A.

Definition 2 A shape is *suitable* if at least one of the points from the set A is on the boundary of this geometrical body (shape).

We present several approaches to find suitable shapes for the finite set.

1.2 *Positive definite symmetric matrix approach (PDSMA)*

Any ellipsoid can be described by a positive definite matrix. Let M^+ be the set of positive definite matrices. In order to find a suitable shape we intend to determine a positive definite symmetric matrix M, such that:

$$\|c - a_i\|_M \leq 1 \qquad \forall i \in \{1, \dots N\},$$

where there exists an $a_j \in A$ such that $\|c - a_j\|_M = 1$, c is the centre of the finite set obtained by an appropriate method, and

$$\|x\|_M^2 = \langle x^T, Mx \rangle.$$

As an approximation of the set A we then consider the object

$$\{x \in \mathbb{R}^n : \|c - x\|_M \leq 1, \}. \tag{1.1}$$

1.3 *Positive diagonal matrix approach (PDMA)*

It is reasonable to only consider the subset $D_+ \subset M^+$ of positive diagonal matrices. In order to find a suitable shape we intend to determine a positive definite diagonal matrix $D = diag(d_1, \dots, d_n)$, such that:

$$\|c - a_i\|_D \leq 1 \qquad \forall i \in \{1, \dots N\}$$

where there exists an $a_j \in A : \|c - a_j\|_D = 1$, c is the centre of the finite set obtained by an appropriate method, and

$$\|x\|_D = \left(\sum_{i=1}^{n} d_i |x_i|^j \right)^{1/j}, \qquad j = 1, 2.$$

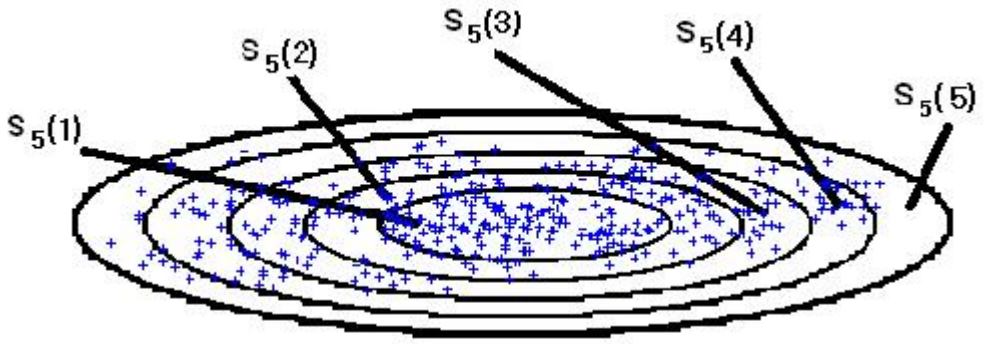

Figure 1: The structure for $L = 5$ for a shape of a finite set

The index j indicates a 1-norm or 2-norm. As an approximation of the set A we then consider the set

$$\{x \in \mathbb{R}^n : \|c - x\|_D \leq 1\}. \tag{1.2}$$

Remark 1 The PDSMA is a generalization of the PDMA for 2-norm, but the dimension of the problem appearing in the PDMA is significantly less.

Once a shape has been found for A, we would like to study the distribution of the points within the shape. Let $L \in \mathbb{N}$. We want to divide the shape into L layers from the center to the boundary, and analyze the distribution of the points within these layers. Figure 1 shows an illustration of the division of an ellipse into 5 layers. The crosses inside are the points of the finite set approximated by the ellipse.

$$\begin{cases} S_L(k) = \left\{ a_i \left| \frac{k-1}{L} \leq \frac{\|a_i - c\|_M}{\max\limits_j \|a_j - c\|_M} < \frac{k}{L} \right. \right\} \text{ for } 1 \leq k \leq L-1 \\ S_L(L) = \left\{ a_i \left| \frac{L-1}{L} \leq \frac{\|a_i - c\|_M}{\max\limits_j \|a_j - c\|_M} \leq 1 \right. \right\} \end{cases}$$

Remark 2 In this structure $card(S_L(L)) \geq 1$, where $card(.)$ represents the cardinality of a set.

Definition 3 The point a_{i_0} is *isolated* if:

$$a_{i_0} \in S_L(L) \qquad and \qquad S_L(L-1) = \emptyset.$$

Very often the set we would like to study consists of two parts. The first part contains a lot of points which are close to each other (the so called *solid part*). The other part contains a few points spaced out around the solid part. The points from the spaced out part could be interpreted as noise. We can use these structures in order to create an algorithm to separate the solid part and the points considered to be noisy.

2 Algorithm

We propose a new algorithm to find an ellipsoidal shape for the finite set of points. This algorithm has three phases.

MAIN ALGORITHM

1. Find the center of the set

2. Eliminate some isolated points,

that are considered as noise

3. Find the ellipsoidal shape

Now we would like to explain each phase more precisely.

2.1 *Find the center*

To find the center of the set A we solve the following optimization problem:

$$\sum_{a\in A} \|x-a\|_2 \to \min \text{ s.t. } x \in \mathbb{R}^n, \tag{2.1}$$

where $\|\|_2$ is the Euclidean norm.

2.2 *Elimination of the isolated points*

A point which is too far from the rest of the points, and will thus induce some noise in the process of finding the shape has to be eliminated from the set A. For that, we need to find a shape in which such a point is isolated. This means that for such a shape, most of the points will be near the center, while the ones to be eliminated will be on the boundary of this ellipsoid.

To find such an ellipsoid, we need the points to be globally the furthest possible from the boundary of the shape, i.e. closer to the center c. This leads naturally to the optimization problem of determining M such that:

$$\sum_{a\in A} \|a-c\|_M \to \min \text{ s.t. } \max_{a\in A} \|a-c\|_M = 1,\ M \in M^+. \tag{2.2}$$

To eliminate the noise, we have to solve the problem (2.2) repeatedly until the noisy data has been removed. Let $L \in \mathbb{N}$, and apply the following algorithm.

ALGORITHM FOR ELIMINATING NOISE

$A_0 = A$
$i = 0$
DO
solve (2.2)
IF $S_L(L-1) = \emptyset$ **THEN** $A_{i+1} = A_i \backslash S_L(L)$
$i = i + 1$
UNTIL $A_i = A_{i-1}$

In our examples, the points were eliminated with both $L = 10$ and $L = 5$ before the next step.

2.3 *Finding the shape*

Once no more noise exists, we can determine the shape. Now we want the boundary of the ellipsoid to be as close to the group of points as possible. We obtain the following optimization problem for M:

$$\sum_{a \in A} \|a - c\|_M \to \max \text{ s.t. } \max_{a \in A} \|a - c\|_M = 1,\ M \in M^+. \tag{2.3}$$

The restriction on the set of positive definite symmetric matrices could be implemented using the Cholesky factorization (for details see [17]). It means that there exists a unique lower triangular $n-$ dimensional matrix G with a positive prime diagonal $g_{i,i} > 0$, $i = 1, \ldots, n$, such that $M = GG^T$. Thus there exists a bijection between the set of positive definite symmetric matrices and the set of lower triangular matrices with positive leading diagonals.

3 Numerical Experiments: comments and descriptions

In order to solve optimization problems, appearing in our algorithm, we used the Discrete Gradient method, proposed and studied in [13] - [15].

3.1 *Subsets*

We would like to present some examples obtained during our experiments with a well known test dataset **Diabete** (see [1] for more information). This dataset contains 2 classes (500 observations in the first class and 268 observations from the second class). All features (1-8) are continuous. After application of a feature selection algorithm, we obtained the subset (1,2,8) of the most informative features. It was possible to draw illustrations in 3-dimensional space, in order to check the efficiency of our algorithm. We present some results for different sets of points. We found 3 clusters in the first class and 3 clusters in the second one (**Sets 1-6**). We also studied each class more precisely (**Sets 7** and **8**).

3.2 *Point elimination*

We present some results obtained by the algorithm for noise elimination. We used the structures of the order $L = 10$ and $L = 5$.

Set	Number of iterations	Points to eliminate	Size of the original set
1	3	119;127;51	137
2	6	156;13;46;80;181;77;85;104;43	226
3	2	19	66
4	2	44	85
5	2	42	137
6	1	none	117
7	1	none	500
8	1	none	268

Table 1: Point elimination for symmetric definite positive matrices

Set	Number of iterations	Points to eliminate	Size of the original set
1	3	119,127	137
2	6	156;13;46;80;181;77;85;104	226
3	2	19	66
4	2	44	85
5	1	none	137
6	1	none	117
7	1	none	500
8	1	none	268

Table 2: Point elimination for diagonal matrices

Remark 3 No actual elimination occurs during the last iteration.

Remark 4 The results for elimination obtained by diagonal and positive definite symmetric positive matrices are slightly different, but in the case of diagonal matrices we have much less variables in the optimization problem. It is therefore reasonable to use diagonal matrices in the elimination process.

Remark 5 If no point is eliminated in a set, the elimination may need to be refined by running the algorithm with another value of L.

3.3 *Shape of the sets*

For each set after point elimination we found a suitable shape and the repartition of the points within this new shape. We present results for the three biggest sets in our example (**Sets 2,7** and **8**).

Set	$S_{10}(1)$	$S_{10}(2)$	$S_{10}(3)$	$S_{10}(4)$	$S_{10}(5)$	$S_{10}(6)$	$S_{10}(7)$	$S_{10}(8)$	$S_{10}(9)$	$S_{10}(10)$
2	5	16	24	23	10	4	3	21	80	31
7	30	112	139	74	42	26	32	19	12	14
8	4	11	35	70	52	31	30	19	8	8

Table 3: The repartition of the points using $S_{10}(k)$ structures, $k = 1, \ldots, 10$.

In the case of the **Set 2**, we observed that many points are close to the boundary of the suitable shape, and therefore the repartition of the points is quite regular. For the **Sets 7** and **8** we couldn't obtain such results. It is possible that we have still some noisy points.

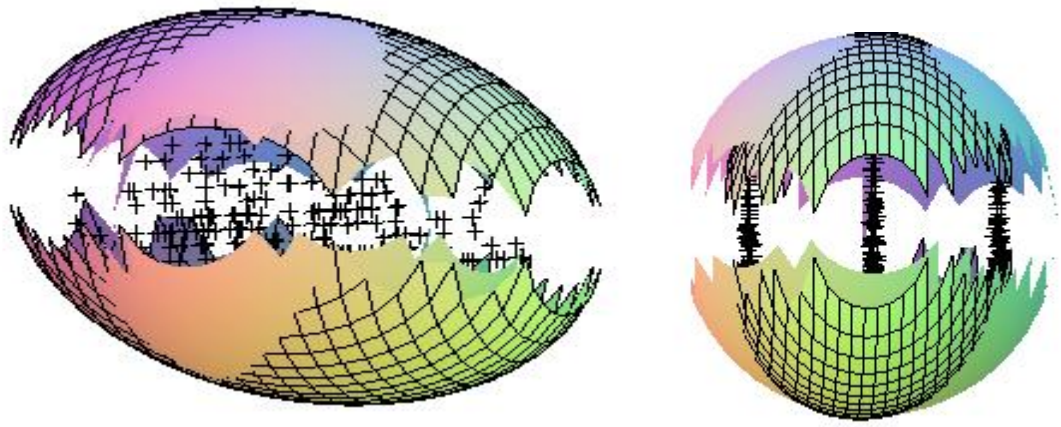

Figure 2: Ellipsoid and ball shapes for **Set 2**

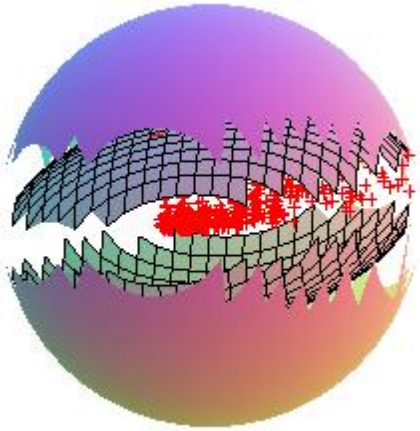

Figure 3: Ellipsoid and ball shapes for **Set 7**

In our computations, as an initial guess, we use the Identity matrix. This means that we start from ball shapes. Figures 2, 3 and 4 illustrate how the shape changed during our computations, the ellipsoid being represented with a grid and the ball without. The set of points is in each case represented by black crosses.

Figure 5 presents two ellipsoids obtained for PDSMA (grid) and PDMA (no grid) for the **Set 2**. The size of the ellipsoid in the case of PDSMA is much smaller. It means that the shape was defined more precisely, but the dimensions of the optimization problem was much bigger.

Remark 6 It is reasonable to consider some intersections of ellipsoids as a suitable shape.

Conclusion

- We developed a new algorithm to approximate a finite set of points. This algorithm allows us to
  - eliminate some noisy points, supposed to be noisy,
  - find a shape as an approximation of a set in the class of ellipsoids.

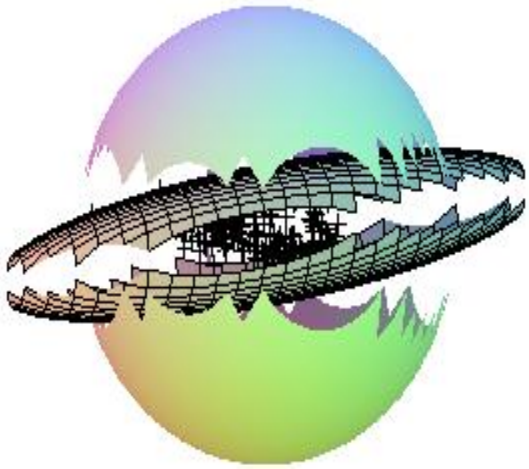

Figure 4: Ellipsoid and ball shapes for **Set 8**

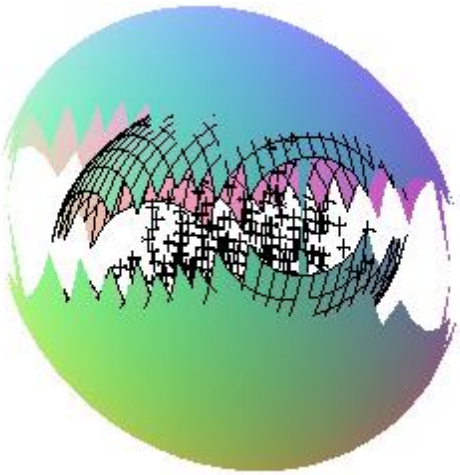

Figure 5: PDSMA and PDMA shapes for **Set 2**

- The optimization problems appearing in the algorithm were solved using Discrete Gradient method.
- The algorithm was used to study sets of points within real world datasets.
- It will be reasonable to continue the research in several directions. Some of them are
  - try to find a suitable shape in other classes of geometrical objects, such as the polyhedra;
  - use and study the intersection of geometrical objects as a suitable shape.

Acknowledgements

We would like to thank Prof. Alex Rubinov and Prof. Adil Bagirov for their support and advice.
This research was supported by the Victorian Partnership for Advanced Computing, Victoria, Australia.
We are also thankful to the anonymous referee for the useful corrections.

References

[1] D. Michie, D. J. Spiegehalter and C. C. Taylor (eds.), *Machine learning, neural and statistical classification*. Ellis horwood series in artificial intelligence, London, 1994.

[2] R. Buyya, G. Mohay, P. Roe (eds.), *Proceedings of the First IEEE/ACM International Symposium on Cluster Computing and the Grid*, 15-18 May 2001, Brisbane, Australia.

[3] J. C. Bezdek, *Fuzzy models and algorithms for patterns recognition and image processing,* Kluwer Academic, Boston: London, 1999.

[4] J. C. Bezdek, *Pattern recognition with with fuzzy objective function algorithms,* Plenum Press, New York, 1981.

[5] D. Dumitrescu, L. C. Jain and B. Lazzerini, *Fuzzy sets and their application to clustering and training,* CRC Press, Boca Raton, FL, 2000.

[6] G. Beliakov *Fuzzy clustering using global optimization*, ICOTA 2002, pp 72-79, Proceedings The 5-th International Conference on Optimization Techniques and Applications, December 15-17, 2001.

[7] A. K. Jain, M. N. Murty and p. J. Flynn *Data Clustering: A Review* ACM Computing Surveys, Vol. 31, No 3, September 1999.

[8] T. Kohonen , *Self-Organizing Map,* 2nd ed. Berlin, Germany, Springer, 1997.

[9] E. W. Forgy, *Cluster Analysis of multivariate data: efficiency versus interpretability of classifications*, Biometrics, 1965, 21, 768.

[10] B. Mirkin, *Mathematical Classification and Clustering*, Kluwer Academic Publishers, 1996.

[11] A. M. Bagirov, A. M. Rubinov and J. Yearwood, *A global optimization approach to classification*, Research Report 01/4, University of Ballarat, http://www.ballarat.edu.au/itms/research_papers.shtml, 2001.

[12] A. M. Bagirov, A. M. Rubinov and J. Yearwood, *A heuristic algorithm for feature selection based on optimization technique*, Research Report 01/2, University of Ballarat, http://www.ballarat.edu.au/itms/research_papers.shtml, 2001.

[13] A. M. Bagirov, *Numerical methods for minimizing quasidifferentiable functions: a survey and comparison*, In: V.F. Demyanov and A.M. Rubinov (eds.), *Quasidifferentiability and Related Topics*, Kluwer Academics Publisher, 33-71, 2000.

[14] A. M. Bagirov, Continuous subdifferential approximations and their applications, University of Ballarat Research Report 01/22 November 2001.

[15] V. F. Demyanov and A. M. Rubinov, *Constructive Nonsmooth Analysis*, Peter Lang, Frankfurt am Main, 1995.

[16] P. M. Murphy, D. W. Aha, *UCI repository of machine learning databases*. Technical report, Department of Information and Computer Science, University of California, Irvine, 1992. http://www1.ics.uci.edu/~mlearn/MLRepository.html.

[17] http://www.cs.ut.ee/~toomas_l/linalg/lin2/node24.html.

[18] A. M. Bagirov, N. V. Soukhoroukova, *Nonsmooth Optimization approach to Data Classification*, Proceedings of the Post-graduate ADFA Conference on Computer Science, 2001.